#4
12-11-99

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

OIPE JC90
OCT 1 2 1999
PATENT & TRADEMARK OFFICE

In re Patent Application of

MEKURIA — Atty. Ref.: 2466-36

Serial No. 09/388,609 — Group: 2741

Filed: September 2, 1999 — Examiner: Unknown

For: A METHOD AND A SYSTEM FOR VOICE DIALLING

* * * * * * * * * * *

Honorable Assistant Commissioner of Patents
Washington, DC 20231

SUBMISSION OF PRIORITY DOCUMENTS

Sir:

It is respectfully requested that this application be given the benefit of the foreign filing date under the provisions of 35 U.S.C. §119 of the following, a certified copy of which is submitted herewith:

Application No.	Country of Origin	Filed
9802990-3	Sweden	4 September 1998

Respectfully submitted,
NIXON & VANDERHYE P.C.

October 12, 1999

By: H. Warren Burnam
H. Warren Burnam, Jr.
Reg. No. 29,366

HWB:lsh
1100 North Glebe Road, 8th Floor
Arlington, VA 22201-4714
Telephone: (703) 816-4000
Facsimile: (703) 816-4100

PRV

PATENT- OCH REGISTRERINGSVERKET

Patentavdelningen

Intyg
Certificate

Härmed intygas att bifogade kopior överensstämmer med de handlingar som ursprungligen ingivits till Patent- och registreringsverket i nedannämnda ansökan.

This is to certify that the annexed is a true copy of the documents as originally filed with the Patent- and Registration Office in connection with the following patent application.

(71) Sökande *Telefonaktiebolaget L M Ericsson, Stockholm SE*
Applicant (s)

(21) Patentansökningsnummer *9802990-3*
Patent application number

(86) Ingivningsdatum *1998-09-04*
Date of filing

Stockholm, 1999-09-28

För Patent- och registreringsverket
For the Patent- and Registration Office

Emma Johnsson

Emma Johnsson

Avgift
Fee 170:-

PATENT- OCH REGISTRERINGSVERKET
SWEDEN

Postadress/Adress
Box 5055
S-102 42 STOCKHOLM

Telefon/Phone
+46 8 782 25 00
Vx 08-782 25 00

Telex
17978
PATOREG S

Telefax
+46 8 666 02 86
08-666 02 86

SS 40117

Telefonaktiebolaget L M Ericsson, ref BT 11848

Inventor: Fisseha Mekuria

A Method and a System for Speech Recognition

TECHNICAL FIELD

The present invention relates to a method and a system for speech recognition, in particular for use when entering commands to a machine, such as a mobile telephone.

BACKGROUND OF THE INVENTION AND PRIOR ART

In speech recognition systems, the main object is to make a machine understand an utterance made by a human speaker. Thus, speech recognition is used for facilitating a man-machine interface (MMI) by means of allowing entering of commands, text and data to the machine directly from the speech.

In speech recognition, the task for the computer is to transform the acoustic input signal into a text, so called transcription. The characteristics of the input signal varies within a broad range for the same word depending on the sex, age, dialect, etc. of the speaker. Furthermore, if several words are entered into the system at the same time, for example if a whole sentence is given to the speech recognition system, the pronunciation of the different words may differ depending on the words preceding and/or succeeding a present word.

Furthermore, the presence of noise and echoing effects may distort the original signal before it enters the speech recognition system.

In general, speech recognition systems can be divided into two main groups:

i) Speaker independent systems and
ii) Speaker dependent systems

Speaker independent systems, in particular those designed for a large vocabulary and for accepting speech without pausing between the different words, i.e. sentences or parts thereof, requires the use of large speech data bases and use different statistical properties of speech and words. Grammatical rules and predictions of what is likely to be said can also be incorporated in such systems.

Speaker dependent systems, on the other hand, and in particular those using a limited vocabulary (typically a few hundred words) and where only one word is spoken at the time, does not require any large data bases. Instead such systems requires training of the particular speaker, or, in some cases speakers, using the system.

A speaker dependent speech recognition system will of course provide a much better performance compared to a speaker independent system for a number of reasons. For example, the number of words is limited and also the system has an exact knowledge of how a particular word should sound, since it has been trained by the particular person using the system.

However, a speaker dependent system can only be used for a limited range of applications. An application, in which a speaker dependent system is to prefer to a speaker independent system is for example entering of commands to a machine.

In such a case the task for the speech recognition system is to transcript the command given orally into a form which can be understood by the machine, i.e. usually a binary word, which is used for controlling the machine. For example, commands such as "Go", "Stop", "Left", "Right" "Yes", "No", etc. can be given orally to a machine, which then executes the corresponding actions.

Nevertheless, even though the number of possible words that the machine has to recognise is limited, typically to a few hundred words, and even though the speech recognition system of the machine has been trained by the voice of the user and therefore has an exact knowledge of how a particular word sounds when spoken by that particular user, a number of possible sources for making a wrong decision still exists.

Thus, noise and echoing effects in the environment will distort the signal entering the speech recognition system. Also, the frequency spectrum of the same word will experience small variations from time to time, and in particular if the speaker

has a cold or the like.

Another problem is that the number of words, even though limited to, typically a few hundred, requires a very large amount of processing power. In a typical speech recognition system the sampling rate is 8000 samples per second and where each sample consists of about 13 bits. This results in that a typical word, which typically lasts for a second, consists of about 100 000 bits.

Thus, in a system where real time constrains exists, for example requiring a response time of 1 second or less, the speech recognition system has to be able to process the large amount of information contained in each word very quickly.

Furthermore, the computational load on the system increases heavily when the number of words increases. This is due to a number of different reasons. Thus, the system has to search a greater number of words when trying to determine which word or command has been spoken. Also, when the number of words/commands increases the risque for that a given command has characteristics which resembles another command increases. In order to avoid a faulty decision the system then has to extract more features from the different words in order to make a correct decision with a required probability. Finally, the possibility that the system interprets a non-existing command word as a command increases if the number of word increases, i. e. the performance of the out of vocabulary rejection (OVR) function decreases.

In a system, which is designed to operate under difficult conditions, such as a mobile telephone comprising a voice controlled dialling system (VCD), i.e. having means for receiving commands orally, and which may be used in a car, the accuracy of existing speech recognition systems is in most cases too low.

Also, US 5515475 describes a speech recognition system designed to build word models starting from phonemes or allophones.

SUMMARY

It is an object of the present invention to overcome some of the problems associated with the prior art and to provide a method and a system having an improved accuracy and which can be used in an environment having echoing effects and which is noisy.

This object and others are obtain by a speech recognition system having its vocabulary arranged in a trellis structure. At each instant only a part of the whole vocabulary of the speech recognition system is searched for a match, depending on where in the trellis structure the speech recognition system is set to search at a particular time. The trellis structure of the vocabulary can preferably be traversed in any suitable way so that, when a certain command is to be given to the system, the system at that time searches the correct part of the vocabulary.

Such an arrangement solves the problem of having to search many words at a time and will significantly increase the accuracy for the speech recognition system, in particular when the speech recognition system is used in a noisy environment and the risque for a wrong decision is thereby reduced.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention will now be described in more detail by way of non-limiting examples and with reference to the accompanying drawings, in which:

- Fig. 1 is a general block diagram of a speech recognition system.
- Fig. 2 is a flow chart illustrating different steps when using a speech recognition system.
- Fig. 3 is a block diagram of a mobile telephone comprising a speech recognition system.

DESCRIPTION OF PREFERRED EMBODIMENTS

In Fig. 1 a general block diagram of a speech recognition system is shown. The system comprises an input terminal 101 connected to an A/D converter 103, which is connected to a signal analysis block 105. The signal analysis block 105 is connected to a block

107 for generating a trellis and to a word recognition block 109. The block 109 is connected to a block 111 for making a decision and which outputs the recognized word. The block 109 is also connected to a vocabulary 113 stored in the speech recognition system, which will be described more in detail below. The vocabulary 113 is also connected to the block 107.

Thus, an audio signal enters the system at the input terminal 101. The analogue audio signal is then converted to a corresponding digital signal in the block 103, for example at a sampling frequency of 8 kHz into digital signal having, for example, 13 bits/sample.

Next, in the block 105 conventional signal analysis of the digitized signal is carried out. The signal analysis may involve feature extraction and other conventional operations such as filtering. The output signal from the block 105 is fed to the block 109 in which the signal is compared with the words stored in the vocabulary 113. The method by means of which the comparison is carried out can be any conventional method, such as pattern recognition or Markov Models.

In the vocabulary the words are arranged in a trellis structure. Thus, at each instant only a part of the whole vocabulary of the speech recognition system is searched for a match. The trellis structure of the vocabulary can be traversed in any suitable way, such as in a tree structure, so that, when a certain command is to be given to the system, the system at that time searches the correct part of the vocabulary. This will solve the problem of having to search many words at a time and will therefore significantly increase the accuracy for the speech recognition system, in particular when the speech recognition system is used in a noisy environment and the risque for a wrong decision is thereby reduced.

In a preferred embodiment the words of the vocabulary are divided into different classes, which in turn can be divided in sub-classes, and so on, as shown in Fig. 1. Thus, first the speaker operating the system selects the class of words he want

to be recognized by the system by means of entering a word corresponding to that class to the speech recognition system. Next, the speaker enters the command word corresponding to the command the speaker wishes to give to the computer system receiving commands from the speech recognition system. If the command word the speaker wishes to enter to the system is located under a sub-class of the class he must, of course, first enter the word corresponding to that particular sub-class.

The entering of commands if preferably assisted by the speech recognition system, for example by means of a voice prompter as described in detail below in conjunction with Fig. 2.

The trellis structure of the vocabulary is advantageous because it significantly reduces the number of word, which the speech recognition system has to search for each entered word or utterance. This improves the accuracy for the speech recognition system, in particular when the speech recognition system is used in a noisy environment and therefore reduces the risque for a wrong decision.

The words in the vocabulary must be entered into the speech recognition system by means of training it. Therefore, when a new word is to be entered into the vocabulary the system is switched into a training mode. This is done by operating the switch S1 between the block 105 and 107.

Thus, when the switch S1 is closed the system is in a training mode. In the training mode conventional training of the system can be carried out. In addition the user of the system can place each word which the system is trained with at any location which he finds suitable in a trellis structure. He can also add classes and sub-classes move words from one location to another or erase words, sub-classes or classes. These operations are preferably assisted by a voice prompter, which can use a speech encoder provided in the system as described below in conjunction with Fig. 3.

In another preferred embodiment the system automatically

generates new classes or sub-classes, when the number of words at a particular location in the vocabulary is higher than a certain pre-set threshold value. A suitable threshold value can be somewhere in the range 20 - 50 words.

An application when the speech recognition system as described herein can be useful is when the speech recognition system is integrated in a mobile phone. In such a case the user of the mobile phone may wish to enter a certain telephone number or wishes that the mobile phone calls a person entered in the phone book of the mobile phone. The vocabulary may then comprise a number of different classes generated during training of the speech recognition system.

In Fig. 2 a flow chart illustrating steps carried out when entering a command to a speech recognition system incorporated into a mobile telephone is shown. Thus, first the speech recognition system of the telephone is switched on in a block 201. Thereupon the speech recognition system waits for the entering of one of the words at the top level of the trellis structure of the vocabulary corresponding to a telephone number, which the user wishes to connect to.

In a preferred embodiment the entering of words can be assisted by a voice propter. Thus if a certain word is entered such as "options" the voice prompter generates all options available at the current position of the trellis. Thus, in this example, when the speech recognition system is incorporated in a telephone and the system is at the top level of the trellis, a typical set of available options could be sub-classes such as "friends", "office" and "family" and also words associated with frequently used numbers and also important numbers can be present at this level, for example "SOS".

If the word corresponding to one of the sub-classes is entered as shown in block 203 the speech recognition system changes position in the trellis structure of the vocabulary. Thus, new options will be available. If, for example, the word "office" is given to the speech recognition systems a first word in the

block 203, the sub-words located under the menu "office" will be searched by the speech recognition system when a new word is entered into the speech recognition system. At the same time the voice propter preferably generates "office" as a confirmation to the user that the speech recognition system has interpreted the word correctly.

Again, if the user wants to know his options he enters "options" and the speech recognition system repeats the options available at this position in the trellis structure of the vocabulary, for example "boss", "secretary" and "up", where "up" corresponds to going up a level in the trellis structure. Other options which may be available are sub-classes at the same level in the trellis, i.e. "friends" and "family" in this case.

If the user enters "secretary" as a second word in a block 205 this word corresponds to a command, in this example a telephone number in the telephone book of the telephone, which the user wants the system to carry out. In a preferred embodiment the voice prompter the repeats "call secretary" and if the user then enters "yes" a command is generated in a block 207 so that the telephone number corresponding to the word secretary in the telephone book is connected to by the telephone and the speech recognition system is switched off as indicated by block 209.

In Fig. 3, a block diagram of a mobile telephone 301 comprising a speech recognition system as described above in conjunction with Fig. 1, is shown. Thus, the mobile telephone 301 has input means 303, which can be a microphone of conventional type connected to an A/D converter 305. The output terminal of the A/D converter 305 is connected to a digital signal processing (DSP) block 307 comprising a speech encoder/decoder (SPE/D) 309 a handsfree signal processing (HFSP) block 311 and an automatic speech recognition (ASR) block 313.

The DSP block 307 is also connected to a micro control (MC) unit 315, to a radio transmission unit 317, comprising a radio transmitter/receiver block 319 and a channel encoder/decoder block 321, to a memory 323 and to a D/A converter 325. The micro

control unit 315 handles all information flow inside the mobile telephone and is set to control the DSP 307 and the radio transmitter/receiver block 319.

The micro control unit 315 is thus also connected to the radio transmission unit 317, which in turn is connected to an antenna 327. The output terminal of the D/A converter 325 is connected to output means 329, such as a loudspeaker of conventional type.

The mobile telephone can then be operated in the manner described above in conjunction with Fig. 2. When switched on the MC unit can automatically set the mobile phone in a mode allowing oral input of commands, and to output instructions via a voice prompter as described above using the speech encoder 309, the D/A converter 325 and the output terminal 329, if required.

Next, when the command(s) has/have been given, MC unit switches off the speech recognition system (ASR) and transmits the telephone number corresponding to the given command via the unit 317 and the antenna 327.

Thereupon, the telephone call is set up using conventional methods and the DSP 307 is set to perform conventional processing, such as acoustic echo cancelling, noise suppression and code the speech efficiently. When the call is terminated the MC unit 315 can again set the DSP unit to receive commands given orally.

The speech recognition system as described herein can be employed in many different computer systems where a high accuracy is required and where many words in the vocabulary also are desired. The trellis structure of the vocabulary makes it possible for the system to cope with these two contradicting wishes. The trellis structure of the vocabulary is independent of the speech recognition algorithm used. In other words it is possible to implement in any large reference vocabulary to be searched. The performance of the out of vocabulary rejection (OVR) function will also increase when using the system as

described herein, since the number of possible entered words is kept to a minimum. The speech recognition system is particularly well suited for, but not limited to, telephone book applications and similar applications.

CLAIMS

1. A speech recognition system comprising a vocabulary, **characterized in** that the words in the vocabulary are arranged in a trellis structure comprising a number of groups of words, so that only one group or a limited number of groups of the entire vocabulary is searched for a word at each time.

2. A system according to claim 1, **characterized in** that the vocabulary is arranged in a tree structure.

3. A system according to any of claims 1 or 2, **characterized by**
- means for outputting the words that the system is set to recognize at a particular moment.

4. A system according to claim 3, **characterized in** that said means is a voice prompter.

5. A system according to any of claims 1 - 4, **characterized by**
- means for automatically generating a new group if the number of words in one group exceeds a certain, pre-set threshold value.

6. A mobile telephone comprising a speech recognition system according to any of claims 1 - 5.

7. A method of speech recognition in a speech recognition system comprising a vocabulary, wherein the words in the vocabulary are arranged in a trellis structure comprising a number of groups of words, **characterized in** that only one group or a limited number of groups of the entire vocabulary is searched for a word at each time.

8. A method according to claim 7, **characterized in** that the vocabulary is arranged in a tree structure.

9. A method according to any of claims 7 or 8, **characterized in** that the available words that the system is set to recognize at a particular moment is output from the system.

10. A method according to claim 9, **characterized in** that the available words are generated by a voice prompter.

11. A method according to any of claims 7 - 10, **characterized in** that a new group automatically is generated if the number of words in one group exceeds a certain, pre-set threshold value.

ABSTRACT

In a speech recognition system the words are organized in a trellis structure. Thus, at each instant the speech recognition system only needs to search a limited part of the entire vocabulary. Such an arrangement solves the problem of having to search many words at a time, which is time consuming and imposes a high computational load on the system, and will therefore significantly increase the accuracy for the speech recognition system, in particular when the speech recognition system is used in a noisy environment and the risque for a wrong decision is thereby reduced.

(Fig. 1)

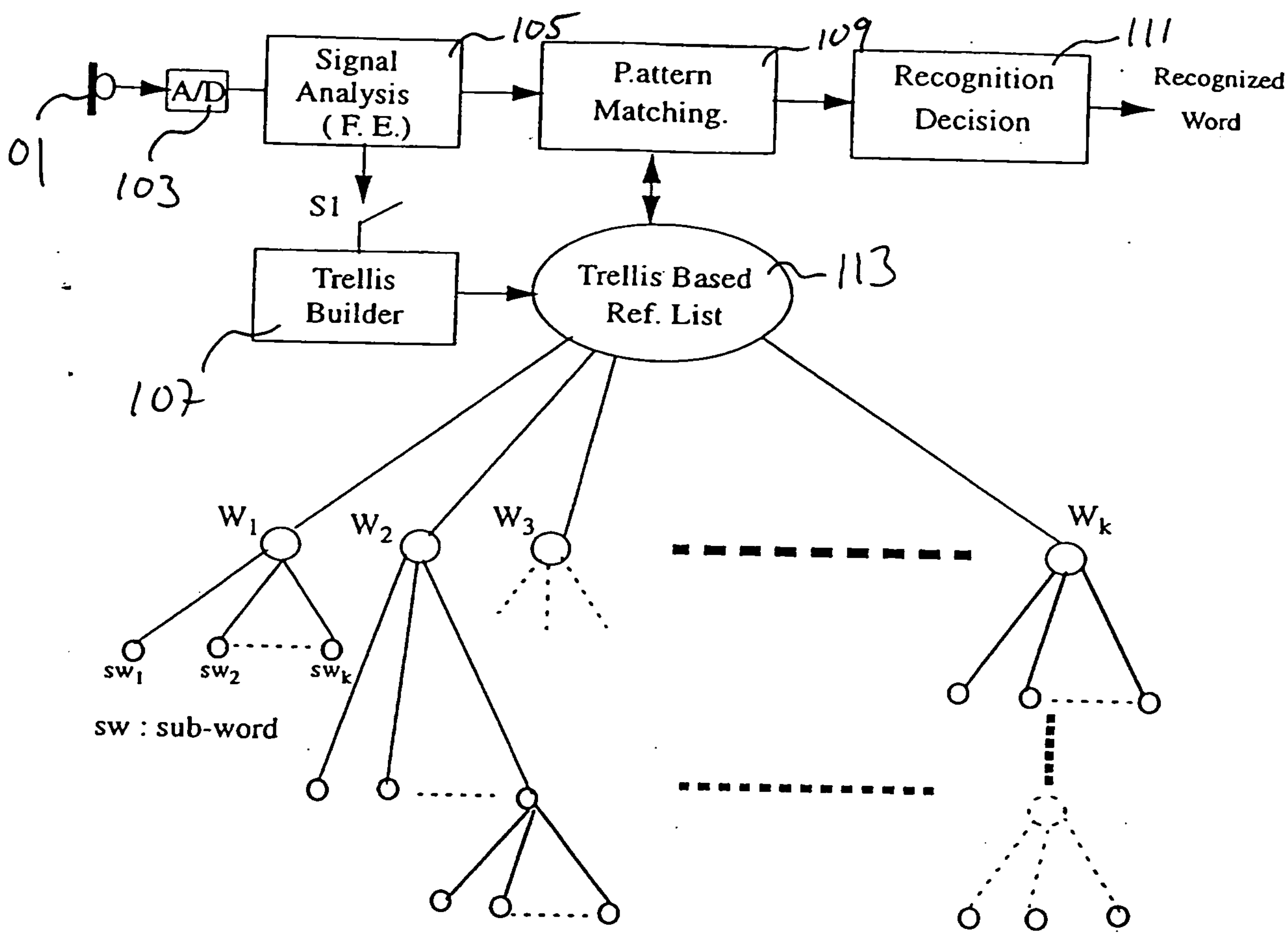

01
A/D
103
Signal
Analysis
(F.E.)
105
S1
Trellis
Builder
107
Pattern
Matching.
109
Recognition
Decision
111
Recognized
Word
Trellis Based
Ref. List
113
W_1
W_2
W_3
W_k
sw_1
sw_2
sw_k
sw : sub-word

Fig. 1.

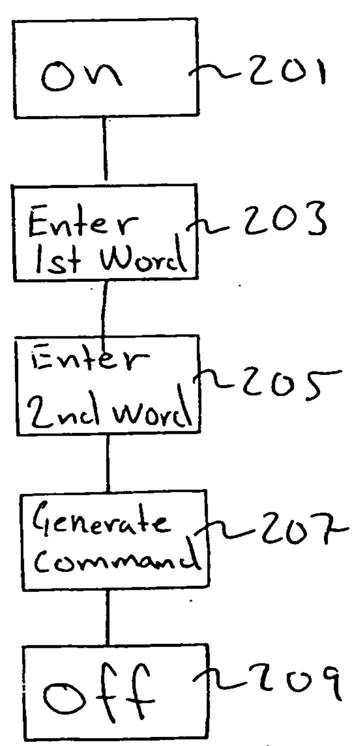
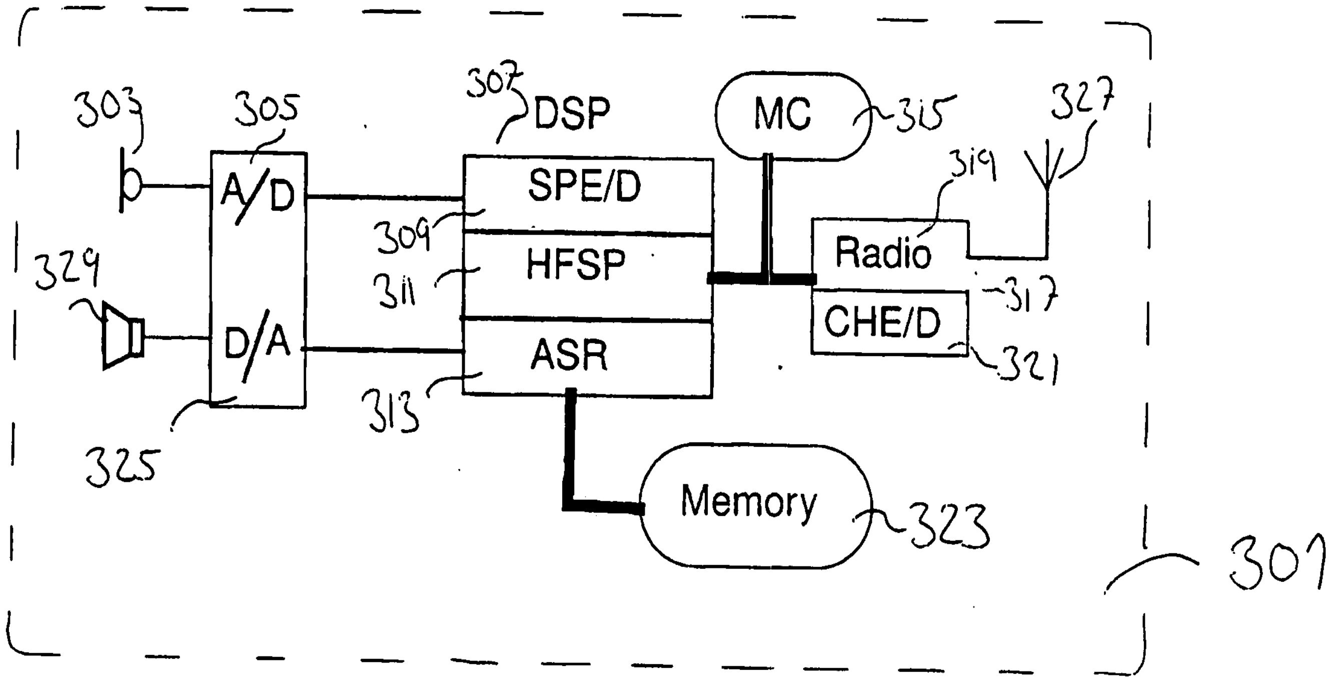

On
201
Enter
1st Word
203
Enter
2nd Word
205
Generate
Command
207
Off
209

Fig. 2



Acronyms, DSP : Digital signal processing,
SPE/D : Speech encoder/decoder
ASR : Automatic speech recognition.
HFSP : Handsfree signal processing, such as acoustic echo cancelling, and noise suppression
CHE/D : Channel encoder/decoder.

Fig.3